DOCUMENT RESUME

ED 408 349 TM 026 749

AUTHOR Morgan, George A.; Gliner, Jeffrey A.
TITLE Helping Students Evaluate the Validity of a Research Study.
PUB DATE Mar 97
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; *Evaluation Methods; Higher Education; Models; *Rating Scales; Reliability; Research Methodology; *Research Reports; Statistical Analysis; Teaching Methods
IDENTIFIERS *External Validity; *Internal Validity

ABSTRACT

Students often have difficulty in evaluating the validity of a study. A conceptually and linguistically meaningful framework for evaluating research studies is proposed that is based on the discussion of internal and external validity of T. D. Cook and D. T. Campbell (1979). The proposal includes six key dimensions, three related to internal validity (instrument reliability and statistics, equivalence of participant characteristics, and control of experience/environment variables) and three related to external validity (operations and instrument validity, population validity, and ecological validity). How to use these scales is illustrated through a study by J. A. Gliner and P. Sample (1996) in which the purpose was to increase the quality of life for people with developmental disabilities. Students have been able to make sophisticated evaluations of studies using rating scales based on these six dimensions, and this method of teaching validity helps students become better consumers of research. (Contains three figures and seven references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
George Morgan
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 408 349

Running head: EVALUATING THE VALIDITY OF RESEARCH

Helping Students Evaluate the Validity of a Research Study

George A. Morgan and Jeffrey A. Gliner

Colorado State University

Presented at the March 1997 AERA Conference in Chicago

BEST COPY AVAILABLE

TM026749

Abstract

Students often have difficulty in evaluating the validity of a study. We propose a conceptually and linguistically more meaningful framework for evaluating research studies based on Cook and Campbell's (1979) discussion of internal and external validity. The proposal includes six key dimensions, three related to internal validity (instrument reliability and statistics, equivalence of participant characteristics, and control of experience/environment variables) and three under external validity (operations and instrument validity, population validity, and ecological validity). Students have been able to make sophisticated evaluations of studies using rating scales based on these six dimensions.

Helping Students Evaluate The Validity of a Research Study

Our students often had difficulty appropriately evaluating the validity of a study. This difficulty was partly due to confusion between different uses of the term validity, for example, between what we call research validity (the validity of the whole study) and instrument validity (the validity of a specific test or measuring instrument). Another source of confusion due to similarity of terms concerns the use of the term selection. Specifically, selection can refer to obtaining a sample from the accessible population, a threat to external validity, or selection can refer to how participants were assigned to groups (e.g., selection threat, Campbell and Stanley, 1966), a threat to internal validity.

The objectives of this paper are to re-examine the topic of research validity, present a revised and relabeled framework, and show how we use it to help students evaluate research (i.e., become better consumers of research). We use much of the conceptualization developed by Cook and Campbell (1979), but our goal is to help students understand the concepts of internal and external validity, by clarifying, without oversimplifying. After using this revised conceptualization of and terminology for research validity, we find that our students learn the concepts better and are more able to apply them correctly when evaluating research studies.

In addition to the confusions mentioned above, there were three issues which arose when our students evaluated a research article using the traditional criteria based on threats to internal and external validity: (a) students tended to assess internal and external validity as "all or nothing" evaluations when we think they should be assessed in a relative fashion, from low to high; (b) some students confused or could not remember the specific threats to internal and external validity because many have peculiar names (e.g., history, interactions with selection, mortality); and (c) students lost track of the main issues because there are so many threats that

deal with very specific, sometimes unusual, situations. Students tended not to see the forest, only the trees.

The Theoretical Framework

To help students understand and remember the important issues related to research validity, our framework maintains the two general headings, internal and external validity, proposed by Campbell and Stanley (1966). Internal validity depends on the strength or soundness of the design and analysis. This definition of internal validity allows us to evaluate non experimental as well as experimental research. Randomized experimental designs are usually to be high on internal validity, but one could judge any study on a continuum from low to high. Campbell and Stanley (1966) said that "External validity asks the question of generalizability: To what populations, settings, treatment variables, and measurement variables can this effect be generalized?"(p. 5).

However, we also build on the reconceptualizations of others. Cook and Campbell (1979) divided internal validity into statistical conclusion validity and internal validity, and they divided external validity into construct validity of putative causes and effects and external validity. Our framework was also influenced by Tuckman (1994) who divided the threats to internal validity into instrumentation bias, participant bias, and experience bias factors. In addition, Smith and Glass (1987) influenced our dimensions; they divided external validity into external validity of operations, population external validity, and ecological external validity. Cook, Campbell and Peracchio (1990) separated statistical conclusion validity and internal validity but acknowledged that they "are alike in that they both promote causal relationships" (p. 514). They also separated construct validity from external validity but state that they "are like each other in dealing with generalizations" (p. 514). Under these four kinds of validity they describe 32 different threats.

Our research evaluation framework includes three dimensions of internal validity (instrument reliability and statistics, equivalence of participant characteristics, and control of experience/environment variables) and three dimensions for external validity (operations and instrument validity, population validity, and ecological validity). We believe that these six dimensions are easier to understand and subsume the many threats to internal and external validity proposed by Cook, Campbell and their colleagues. Before discussing these dimensions of internal and external validity, we would like to put them in the broader context of reliability and validity. Doing this has helped our students avoid some of the confusion mentioned above.

Research vs. Instrument Reliability and Validity

It is important to distinguish between merit or worth of the whole study (research validity) as opposed to the quality of an instrument or test used in a study (instrument validity). For example, Krathwohl (1993) criticized Campbell and Stanley's (1966) use of the term validity to assess research because, "…there are other contexts where the term validity is also appropriately used" (p. 270). Figure 1 shows that instrument reliability and validity (upper half of the figure) are different from, but related to, research reliability and validity (lower half), and it also shows how both fit into an overall conception of reliability and validity. While we recognize the importance of reliability issues, our major aim with the present paper is to clarify validity issues which seem to cause so many problems for students. The figure, accompanied by definitions and examples of each term, helps students learn and understand the difference between instrument validity and research validity. We believe that Figure 1, by showing where the various aspects or types of validity fit, helps prevent confusion among test validity, construct validity (of a test), internal validity, and external validity. Our labels for the first (a) and fourth (d) dimensions of research validity show how instrument reliability and validity affect internal

and external validity.[1] We will now turn to a discussion of the six dimensions of research validity, including some of the examples that we give students to help them evaluate each dimension.

Evaluating Internal Validity

A good study should have moderate to high internal validity on each of the following three dimensions. If not, the author should be, at the very least, cautious not to say that the independent variables influenced, impacted, or caused the dependent variables to change. We ask students to rate the study as a whole from low to high on each of the three scales and to explain why they made each rating. Figure 2 shows these rating scales and several issues designed to help students discuss each of the aspects of internal validity.

Instrument reliability and statistics. Although Cook and Campbell (1979) called our first dimension of internal validity statistical conclusion validity, we have modified the name to emphasize the importance of instrument reliability and to indicate where it fits into the overall framework of Figure 1. There are four important issues that the student should consider when evaluating research and making the overall rating of instrument reliability and statistics.

The first issue is whether the variables are measured reliably. Students are asked here to consider the overall reliability of the instruments, as a group. Although instrument reliability influences power, we think it should be singled out. A principle of measurement is that a test can not be valid if it is not reliable. Analogously, a study's research validity is reduced if one or more key variables, including the independent variable, is unreliable.

---

[1] We realize that instrument validity affects internal as well as external validity, but in order to keep this framework relatively simple and straightforward, we have located instrument validity in the fourth dimension, operations and instrument validity. This is consistent with Cook and Campbell (1979) who put their category of construct validity of putative causes and effects under external validity.

Second, does the study have appropriate power? Although there are other methods of increasing power, such as decreasing variability or increasing the strength and consistency of the independent variable, we have focused on whether there are an appropriate number of participants. Students intuitively know that small samples cause problems, but Cook and Campbell (1979) raise the issue of having too much power, especially when an exceptionally large sample size yields a statistically significant, but very weak, relationship. Reporting effect size is one way to deal with this issue.

The third issue involves the selection of the proper statistical method. Although researchers sometimes select the wrong statistic, Cook and Campbell (1979) pointed out that more often problems involve violating assumptions of the statistical tests or making multiple tests without adjusting the alpha level. The fourth issue involves the proper interpretation of the statistical analysis. For example, a statistically significant result from a single factor ANOVA with three or more levels does not imply that all levels or groups are different from each other.

Equivalence of participant characteristics. A key question is whether groups that are compared are equivalent in all respects other than the independent variable or variables. Campbell and Stanley (1966) described a number of specific threats to internal validity, several of which (selection, statistical regression, experimental mortality, and various interactions) are subject factors that could lead to a lack of equivalence of the participants in the groups and, thus, influence the results. Our students often found the labels of Campbell and Stanley's threats confusing, and we find their categories more complex than necessary for a basic understanding of internal validity. Another problem with the emphasis on threats to internal validity is that a threat often only instructs the student about why the groups might not be equivalent, but does not get at how to correct the problem.

Experimental research achieves equivalence through random assignment of participants to adequately sized groups. However, in non-experimental research, random assignment to groups is not accomplished. Other methods such as random assignment of treatments to intact groups, analysis of covariance, matching, or checking for pretest equality of groups after the fact are attempts to achieve equivalence. Figure 2 shows that students are asked to rate this dimension high (random assignment to groups), medium (some attempts to equate groups), or low (unequal groups, perhaps due to self assignment, with little attempt to make them more nearly equal).

Control of Experience/Environment Variables. We group several of Campbell and Stanley's (1966) other threats to internal validity under a category that deals with the effects of extraneous experiences and/or environmental conditions during the study. Cook and Campbell (1979) address this problem when discussing threats to internal validity that random assignment does not eliminate (see p. 56). Some of these threats occur because participants gain information about the purpose of the study while it is taking place. The first issue is whether extraneous variables or events affect one group more than the other. For example, if participants learn that they are in a control group, they may not try as hard, exaggerating differences between the intervention and control groups. Alternately participants in the control group may overcompensate, eliminating differences between the two groups. One method to prevent some of these external influences is to isolate the intervention group from the control group. For example, when performing research in a school system, it might be good to have the intervention and control groups from different schools. A second method to reduce extraneous influences, especially in exploratory studies, is to shorten the time of the intervention in order to lessen the chances for external variables to have an effect on one group or the other. A third method that might be used to try to eliminate external influences is to increase the potency of the treatment.

Large differences between the treatment and the control are less likely to be obscured by extraneous variables.

A second issue is whether something other than the independent variable affects both or all groups. Historical events or maturation could have an effect on the dependent variable that is mistaken for an effect of the independent variable. This problem often occurs when one is comparing the effects of two different treatment groups. When such a study is performed, and no statistically significant differences are obtained between the treatment groups, the author may conclude that both treatments worked equally well. However, if a control group which did not receive the treatment was not included, it could be that neither treatment worked and that both groups got better over time for some other reason. A "no treatment" control group may not have been included because of ethical considerations. If a control group cannot be included, then the authors need to find a way to document that participants would not have improved over time. This might be done by citing previous research in this area. Another method to deal with this situation is where treatment is delayed, for at least one of the treatment groups, so that the effects of no treatment could be assessed.

Evaluating External Validity

A good study should be rated high on all three aspects of external validity, or, at least, authors should be cautious about generalizing the findings to other measures, populations, and settings. Figure 3 shows the three rating scales for external validity and several issues intended to draw attention to issues that need to be considered in making an overall rating.

Operations and instrument validity. This category has to do with whether the variables, including any interventions, are appropriately measured/defined and are representative of the concepts or constructs under investigation. Research articles deal with this question, in part, as

test or instrument validity. We ask students to make an overall judgment of the validity and generalizability of the operational definitions of the several key variables in the study. Unfortunately, sometimes these operational definitions could be interpreted as measuring some other construct, and, thus, results will not generalize to studies that use other instruments.

There is obvious similarity between this dimension and construct validity of cause and effect (Cook et al., 1990), but we have found that our terminology is clearer for students. Nevertheless, we have found that this dimension has proven to be the most difficult one for students to understand for at least two reasons. One problem is that students tend to examine only the dependent variable(s), perhaps due to the fact that researchers are more inclined to discuss validity of the dependent variable. The other problem is that the independent variable, especially in experimental research often does not represent an obvious theoretical construct. Continued emphasis on these two points with concrete examples is highly recommended.

Cook and Campbell (1979) provided an excellent example of this issue using the construct of supervision. Supervision was operationally defined as the supervisor being ten feet or less away from the worker. However, the operational definition could also be relevant to the construct of stress because having a supervisor that close could lead to increased stress among workers. Therefore, are the investigators assessing the effects of supervision or the effects of stress?

Population Validity. This second aspect of external validity involves the participants of the study and how they were selected. The issue of population external validity is more complex than an evaluation of whether a probability (e.g., random) sample was selected from the accessible population. The real question is whether the actual sample of participants represents the theoretical or target population. We have found it helpful to ask students to identify the (a)

apparent theoretical population (which is usually not specified in a research article), (b) the accessible population, (c) the sampling design/method, (d) the selected sample and (e) the actual sample of participants who completed participation in the study. Doing this helps our student think about the representativeness of b, d and e as indicated on the rating form. It is possible that the researcher could use a random or other probability sampling design but have an actual sample that is not representative of the theoretical population. This could be due to a low response rate or to the accessible population not being representative of the theoretical population. The latter problem seems almost universal in behavioral sciences, in part due to funding and travel limitations. Except in national surveys, researchers commonly start with an accessible population from the local school district, community, clinic or animal colony.

Ecological Validity. The third dimension of external validity deals with whether the conditions/settings, times, testers, and/or procedures are natural, and, thus, whether the results can be generalized to real life outcomes. Field research is more likely to be high on ecological external validity than laboratory procedures, which are usually artificial. We regard most of the self-report measures, especially questionnaires, to be somewhat artificial because they do not directly measure the participant's actual behavior in a typical environment. Most of our students appear to grasp this dimension of external validity. The exceptions usually involve a teaching method or a therapeutic technique that, while representative of the construct, has not been carried out in a similar setting or for the same duration as the actual or proposed treatment. For high ecological validity a treatment should be conducted in a realistic setting, by an appropriate therapist, in a manner and for a time period that is similar to that which is how it is usually given. Cook and Campbell (1979) also include generalization to past and future times under external

validity. Thus, students should consider here whether a study is likely to be bound to a specific time in history.

The Relative Importance of Different Validity Categories

We instruct our students that it is nearly impossible for a single study to achieve high ratings for each of the six dimensions of validity. Typically, researchers sacrifice one dimension to strengthen another dimension. For example, studies performed in tightly controlled situations, such as a laboratory, usually sacrifice one or more measures of external validity in favor of strong internal validity. On the other hand, studies performed "in the field" usually surrender some degree of internal validity, in favor of strong ecological external validity.

Should a study be judged more harshly if it is weaker on certain validity dimensions than on others? Campbell and Stanley (1966) commented that:

> Both types of criteria (internal and external validity) are obviously important, even though they are frequently at odds in that features increasing one may jeopardize the other. While internal validity is the sine qua non, and while the question of external validity, like the question of inductive inference, is never completely answerable, the selection of designs strong in both types of validity is obviously our ideal (p. 5).

Cook and Campbell (1979) also addressed the issue in some depth. They suggested that if one is interested in testing a theory, then internal validity and construct (operations) validity appear to have the highest priority. Obviously, the constructs used in the study must represent those in the theory. Also, one would need to show a causal relationship between or among variables when testing a theory. On the other hand, Cook and Campbell point out that if one is to

perform applied research, then more emphasis should be placed on external validity, especially if the research involves specific diagnostic groups.

An Example of How Students Use the Framework

To illustrate how students use this framework and Figures 2 and 3 to evaluate research, consider a study by Gliner and Sample (1996). The purpose of the study was to increase quality of life for persons with developmental disabilities who were employed in sheltered work or supported employment, using an intervention of community life options. The study attempted to achieve high internal validity by randomly assigning participants to either a community life options intervention or to their present situation. The study also attempted to achieve high external validity by carrying out the conditions in the actual setting. However, obtaining good research validity on all six dimensions could not be accomplished.

Internal validity. The instrument reliability and statistics were judged overall to be medium high. Reliability of the measures was good. However, statistical power was constrained because there were a limited number of persons who fit the criteria to be in the study (persons with developmental disabilities who were employed in supported or sheltered work). Thus, the ability to detect a relationship was reduced. The choice of statistics and their interpretation was judged to be high. Equivalence of subject characteristics was rated high because participants were randomly assigned to intervention conditions. However, a cautionary note could be raised because random assignment of participants to conditions may not make the groups equivalent with small numbers. Control of experiences and the environment, was constrained by an emphasis on ecological validity so it was judged to be medium to low. In a community setting, where choice was experienced differently by different participants, it was difficult to insure that the experiences of each group were not influenced by outside variables.

External validity. In terms of operations and instrument validity, the dependent variable, quality of life for persons with developmental disabilities, had been used several times with this population to measure quality of life among individuals who had moved out of institutionalized settings into community settings. However, the instrument may not have been appropriate for measuring changes following intervention in only one life area. In addition, the instrument may have been intended for lower functioning participants. In terms of the independent variable, the intervention seems appropriately named and generalizable. Overall this dimension was rated medium. Population external validity was considered to be medium low because the sample was limited to persons in one city, and there was not a random selection of participants even from that city. Instead, the sample was one of convenience. Thus, the accessible population might not represent all persons with developmental disabilities. Because the intervention was a real one and took place in an actual community setting, ecological external validity was judged to be high.

Conclusions

The present framework divides research validity into six main dimensions, each of which can be rated from low to high. We provide students with a detailed rating sheet (Figures 2 and 3) and with many examples that elaborate on how to use the dimensions to evaluate any study. This method of teaching research validity is important because we believe that it helps students become better consumers of research than when we used the more traditional methods to evaluate research. Our paper is not is not intended to be a report of a quantitative, data-based study. However, we have informally collected qualitative evidence in the form of student evaluations and exam answers which support the contention that there is less confusion and enhanced ability to evaluate research appropriately.

This conceptualization should be helpful to teachers of a range of psychology and other behavioral science courses because it presents what we believe is an improved framework for teaching how to evaluate research, a goal common to many such courses. In basic courses, like introductory psychology, the figures may have to be simplified. Perhaps, the six dimensions could be reduced to the two basic dimensions of internal and external validity with three or four key issues (e.g., instrument reliability, power, random assignment and control of extraneous variables) for each. We also believe that Figure 1 will help students at all levels understand how the concepts of test or instrument reliability and validity fit into the framework for evaluating whole studies using the concepts of internal and external validity. We have found that, with this figure students are less likely to be confused than previously when they thought there were several unrelated uses of the term validity.

References

Campbell, D. T., & Stanley, J. C. (1966). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Boston: Houghton Mifflin.

Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi experimentation. In M. Dunnette & L. Hough (Eds.), Handbook of industrial and organizational psychology (pp. 223-326). Chicago: Rand McNally.

Gliner, J. A., & Sample, P. (1996). A multimethod approach to evaluate transition into community life. Evaluation and Program Planning, 19, 225-233.

Krathwohl, D. R. (1993). Methods of educational and social science research. New York: Longman.

Smith, M. L., & Glass, G. V. (1987). Research and evaluation in education and the social sciences. Englewood Cliffs, NJ: Prentice-Hall.

Tuckman, B. W. (1994). Conducting educational research (4th ed.). Fort Worth, TX: Harcourt Brace.

RELIABILITY Stability or Consistency	**VALIDITY** Accuracy and Representativeness
Instrument (or Test) Reliability The participant gets the same or a very similar score from a *test*, *observation*, *rating*, etc. Types of *test reliability*: a. Test - Retest Reliability b. Equivalent Forms Reliability c. Internal Consistency Reliability d. Interrater Reliability	**Instrument (or Test) Validity** The score accurately reflects/measures what it was designed or intended to measure. Types of *test validity*: a. Face Validity b. Content Validity c. Criterion-related Validity d. Construct Validity
Research Reliability If repeated, the *study* would produce similar results. This is called *replication*.	**Research Validity** The results of the study are accurate and generalizable. Dimensions of the *validity of a study*: **Internal Validity** (accuracy, causality) a. *Instrument Reliability* & Statistics b. Equivalence of Participant Characteristics c. Control of Experience/Environment Variables **External Validity** (generalizability) d. Operations & *Instrument Validity* e. Population Validity f. Ecological Validity

Figure 1. Similarities and differences between instrument and research reliability and validity

INTERNAL VALIDITY/CAUSALITY

a) Instrument Reliability and Statistics

LOW — *MEDIUM* — *HIGH*

Low on all — Mix on below — High on all

Base overall rating on:

(Low vs. High unless otherwise noted)

1) Reliability of instruments/ measures
2) Appropriateness of power (e.g., too few vs. too many participants)
3) Appropriateness of statistical techniques used
4) Appropriateness of interpretation of the statistical analysis

b) Equivalence of Participant Characteristics

LOW — *MEDIUM* — *HIGH*

Groups very different No control of subject characteristics — Some attempts to equate groups — Random assignment to groups

Base overall rating on:

1) Equivalence of the groups on attributes other than the independent variable
2) Attempts to equate participant characteristics through matching, ANCOVA, checking pretest, etc.

c) Control of Experiences and Environment Variables

LOW — *MEDIUM* — *HIGH*

Extraneous variables not controlled — Attempts to control experiences/environment — All extraneous variables are controlled, eliminated or balanced.

Base overall rating on:

1) Extent to which extraneous variables/events could affect one or both groups and obscure true effect, if any, of the independent variable.
2) Attempts to reduce extraneous influences.

Figure 2. Rating Scales to evaluate the internal validity of the findings of a study

EXTERNAL VALIDITY/GENERALIZABILITY

d) Operations and Instrument Validity

LOW	*MEDIUM*	*HIGH*
Treatments and measures not valid or generalizable	Some problems with validity or generalizability	Treatments and measures are valid and generalizable

Base overall rating on:
(Low vs. high unless otherwise noted)

1) Operational definitions of the treatment are appropriate to the concept of interest
2) Validity and Generalizability of the dependent variable measures

e) Population

LOW	*MEDIUM*	*HIGH*
Actual sample unrepresentative	Some attempt to obtain a good sample	Actual sample representative of theoretical population

Base overall rating on:

1) Representativeness of accessible population from theoretical population
2) Sampling method from accessible population (non probability vs. probability)
3) Return/response rate

f) Ecological

LOW	*MEDIUM*	*HIGH*
Setting, tester, time, and procedures unnatural	Somewhat artificial; e.g. questionnaire	Setting , tester, time, and procedures natural

Base overall rating:

1) Naturalness of setting/conditions (lab vs. field)
2) Rapport with testers/observers.
3) Naturalness of procedures/tasks
4) Appropriateness of timing and length of treatment
5) Extent to which results are restricted to one time in history

Figure 3. Rating Scales to evaluate the external validity of the findings of a study